Variable Selection with Spatially Autoregressive Errors: A Generalized Moments LASSO estimator

Variable Selection with Spatially Autoregressive Errors: A Generalized Moments LASSO Estimator

Liqian Cai[+ *] Arnab Bhattacharjee[#] Roger Calantone[$]
Taps Maiti[+]

February 5, 2019

Abstract

We propose generalized moments LASSO estimator, combining LASSO with GMM, for penalized variable selection and estimation under the spatial error model with spatially autoregressive errors. We establish parameter consistency and selection sign consistency of the proposed estimator in the low dimensional setting when the parameter dimension $p <$ sample size n, as well as the high dimensional setting with p greater than and growing with n. Finite sample performance of the method is examined by simulation, compared against the LASSO for IID data.

*Correspondence: Liqian Cai, Department of Statistics and Probability, Michigan State University, East Lansing MI, USA. E-mail: cailiqia@stt.msu.edu. [+] Department of Statistics and Probability, Michigan State University, East Lansing MI, USA. [$] Department of Marketing, Michigan State University, East Lansing MI, USA. [#] Spatial Economics and Econometrics Centre, Heriot-Watt University, UK. We thank organisers and participants in seminars at the University of Illinois and Indian Statistical Institute, the USC Dornsife INET Conference on Big Data in Economics, and invited presentations at the 26$^{\text{th}}$ (EC)^2 Conference and American Statistical Association JSM (Business & Economic Statistics Section) for valuable comments and suggestions. The usual disclaimer applies.

The methods are applied to estimation of a spatial Durbin model for the Aveiro housing market (Portugal).

Key words: LASSO; GMM; Spatial autoregressive errors; Hedonic house price models.

JEL Classification: C32; C52; R31.

1 Introduction

This paper develops methods for estimation and inferences on spatial regression models in high dimensional settings, where the number of parameters is large, potentially even larger than the sample size. In recent years, there has been a proliferation of data-rich environments in different walks of science, society and the economy where such high dimensional applications arise. Today's economists, scientists and policy makers have access to roughly a thousand times more data, and a thousand times more powerful analytic and computational tools, than would have been available for making the same inferences as recently as the early 1990s. Data from high-frequency economic transactions, detailed macroeconomic data collected by a multitude of sources with varying data quality and varying sampling frequencies, and data on large economic and social networks are just a few examples of the content of enormous databases that are now subject to thorough examination. As the Big Data wave washes over them, decision makers find themselves acquiring new sources of data, greater volumes of data and more real-time context. Thus, many applied contexts have emerged where methods for variable selection in large dimensional and dependent data setting are very useful; see, for example, Castle and Hendry (2014), Varian (2014), Belloni et al. (2016) and Chudik et al. (2016).

Concurrently, with the growth of computer capabilities, databases are becoming progressively larger and more complex, making traditional statistical methods less effective or sometimes even unsuitable. In particular, understanding the role of information structure on improved inferences and decision making has turned out to be a key issue in high dimensional and big data contexts. Together, there have been significant developments

in the literature on high-dimensional data sciences based on statistics and econometrics, of methods to estimate regression models with a large number of potential regressors, but where only a few of the parameters are non-zero, that is, where the true model is sparse. In particular, a lot of attention has been devoted to penalized estimators of which the most popular is the LASSO (least absolute shrinkage and selection operator), proposed originally by Tibshirani (1996), and subsequently with many enhancements and variants which Varian (2014) termed "LASSO and friends". However, almost the entirety of this literature assumes independent sampling, where the errors are assumed to be independently and identically distributed (IID). This assumption is not valid in many situations, and certainly not in applications where the data are spatial, which is the domain of the current paper. We develop the LASSO in a spatial regression context and apply the proposed methods to the study of hedonic house price models.

Specifically, in this paper, we propose generalized moments LASSO as a two-stage estimator, combining a GMM estimator to account for spatial dependence, along the lines of Kelejian and Prucha (1999), with the LASSO for high dimensional variable selection. Variable selection methods such as the LASSO, developed under the independent error assumption, do not perform well for dependent data in general (Kock and Callot, 2015) and spatial regression models in particular (Nandy et al., 2016). Our proposed methods relate to a spatial regression model, first proposed by Cliff and Ord (1973) and Anselin (1988), where spatial autoregressive (SAR) errors capture the spatial autocorrelation of errors across the cross-section units. The model and its estimation using GMM have been found useful in many large dimensional contexts; see for example, Brady (2011) and Flores-Lagunes and Schier (2012). Here we extend the context to the situation when there is, in addition, a model selection problem with a large number of potential regressors; our application and LASSO-based model selection methodology is related to Nowak and Smith (2017), but our spatial Durbin model is similar to Lee and Yu (2016).

The remainder of the paper is organized as follows. In Section 2, we provide a brief overview of the background literature and context and place our contributions within that context. Section 3 introduces our generalized moments LASSO, where we combine the

GMM estimator with the LASSO estimator in order to select and estimate the nonzero components of the regression parameter in a dependent (spatial autoregressive) error setting. Section 4 develops the asymptotic properties of the estimator which includes parameter consistency and model sign consistency when the dimension of parameter p is fixed and smaller than the sample size n. Then, Section 5 extends the asymptotic properties of the estimator to the high dimensional setting when p can be increasing with n. Section 6 reports on a simulation study of the performance of the estimator for different choices of parameter ρ in the spatial autoregressive error model with appropriate selection of the penalty parameter λ_n, which we discuss later. Section 7 illustrates the proposed methods by application to the estimation of a spatial Durbin hedonic house price model. Section 8 concludes. The proofs of lemmas and theorems are collected in an Appendix (included as a Supplement).

2 Literature

As discussed above, the work in this paper builds upon the LASSO and related methods, "LASSO and friends" (Varian, 2014).[1] While the LASSO has been applied widely in many fields, most of the available theoretical results relate to an IID setting. This is not useful in the spatial context of this paper. In this section, we discuss briefly some other developments in the literature, relating either to the LASSO for dependent data or other related methods. Likewise, we briefly discuss GMM estimation methods for spatial regression models but in settings where model selection is not an object of inference. Then, this paper develops a method where the GMM is combined with the LASSO to develop a generalized moments LASSO estimator.

Alternate approaches to the LASSO have also been proposed for model selection, the most prominent of which are Bayesian model averaging and spike-and-slab methods; see Ishwaran and Rao (2005) for an initial and highly influential contribution, and Varian

[1] See Bühlmann and van de Geer (2011), Belloni and Chernozhukov (2011) and Kock and Callot (2015) for reviews and references.

(2014), Cuaresma and Feldkircher (2013) and Nowak and Smith (2017) for recent applications in similar contexts to this paper. Much effort has been devoted to establish oracle property of the above methods, that is, establish conditions under which the procedure correctly detects the sparsity pattern, placing zeroes precisely on variables that have zero coefficient values, and furthermore, that the estimates of the non-zero parameters are consistent and asymptotically efficient.

The above methods were developed mainly in the statistics literature. Concurrently, research in econometrics has provided new methods and insights that relate more specifically to economic and financial applications, including some dependent data contexts. The Indicator Saturation methods (Impulse Indicator Saturation, IIS, and Step Indicator saturation, SIS) of Hendry et al. (2008) and Castle et al. (2015) relate to general-to-specific modelling context where a large number of indicator variables are included to allow for arbitrary outliers and potential omitted variable bias. Statistical properties of these methods were initially developed for IID data, but later generalized to dynamic regression models (possibly with unit roots) and structural breaks; see, for example, Johansen and Nielsen (2009) and Castle and Hendry (2014). Likewise, the factor model has become extremely popular in empirical work since Stock and Watson (2002), and this was extended to a factor-augmented global vector autoregression model (FAVAR and GVAR) initially by Pesaran et al. (2004). Then, this model was taken to infinite dimensional VARs in Chudik and Pesaran (2011) and to a spatial strong dependence setting by Chudik et al. (2016). Oracle properties for the adaptive LASSO (Zou, 2006) in this model was recently established by Kock and Callot (2015).

A third line of the econometrics literature has focussed on Generalized Method of Moments (GMM) in a large-dimensional setting. LASSO-based methods here have focussed either on a large number of potential instruments (Belloni et al., 2012), or its variants, importantly the square root LASSO of Belloni et al. (2011), where the optimal penalty does not require a plug-in estimator of the error variance; and most recently to construction of an instrument that mitigates against model selection errors (Belloni et al., 2016). In a related setting, Caner and Zhang (2014) extended a variant of the LASSO,

the adaptive elastic net (Zou and Zhang, 2009), to the GMM method. The current paper also develops results for the LASSO in a GMM context, but applied to spatial regression models, following from the approach proposed by Kelejian and Prucha (1999, 2010), Kapoor et al. (2007) and Lin and Lee (2010).

Our paper shares its spatial context with some other developments in the recent literature, while being distinctly different in its objective and approach. Cuaresma and Feldkircher (2013) developed a Bayesian model averaging method that uses spatial filtering in order to account for uncertainty in spatial weights. In a social network setting, Flores-Lagunes and Schiner (2012) take the workhorse spatial regression models to a binary choice rational expectations context with sample selection. Bhattacharjee et al. (2016) developed a method based on functional regression model that allows unrestricted spatial heterogeneity and endogeneity of spatial weights. In a panel data context, Bailey et al. (2016) developed a two-stage approach for spatio-temporal modeling where the spatial weights matrix is not specified *a priori*; for an alternate method in a similar context, see Ando and Bai (2016). Note that, Chudik et al. (2016) also model spatial strong dependence in the infinite dimensional VAR context using a factor structure, and this is similar to Ando and Bai (2016) and Bailey et al. (2016). Unlike the current paper, the above literature does not address issues of model selection. However, while Bailey et al. (2016) model spatial weak dependence using multiple tests on spatial autocorrelations, LASSO provides an alternate approach. A LASSO-based method for estimating the spatial weights matrix in an endogenous spatial lag model with spatial heterogeneity is proposed in Ahrens and Bhattacharjee (2015); see also Lam and Souza (2016). However, the above spatial literature does not consider model selection in a high dimensional context, and neither has the model selection literature accounted for spatial dependence in any substantial way.[2]

Our methods apply to a model with spatially autoregressive errors first proposed by Cliff and Ord (1973). This model extends autocorrelated errors in the time series context to the spatial dimension and is a variant of the model suggested in Whittle (1954). In

[2]See Nandy et al. (2016) and Feng et al. (2016) for related research.

this spatial model, the error term corresponding to a cross-section unit is modeled as a weighted linear combination of errors corresponding to other cross-sectional units, plus an idiosyncratic error. To be precise, the error u_n is generated as

$$u_n = \rho M_n u_n + \epsilon_n,$$

and the underlying regression model with SAR errors u_n is specified as

$$Y_n = X_n \beta + u_n.$$

The term $M_n u_n$ is often referred to as the "spatial lag". Typically the idiosyncratic errors ϵ_n are assumed to be IID with mean zero and variance σ^2 and the parameters of interest are ρ, σ^2 and β. We assume that the $n \times n$ spatial weight matrix M_n is known. Contrary to time-series models which are associated with unidirectional time flow, spatial data can be viewed as multidirectional, with each location correlating with all the other nearby locations in every direction; thus there is no natural ordering of the data. Because of this particular characteristic of spatial processes, a simple transposition of time-series methodologies cannot be applied.

Estimation of the above model with spatially correlated errors proceeds either by the use of parametric methods based on maximum likelihood (ML) or quasi maximum likelihood (QML) (Anselin, 1988; Lee, 2004; Yu et al., 2008; Lee and Yu, 2010) or the GMM approach proposed by Kelejian and Prucha (1999, 2010) and Lin and Lee (2010). A major limitation of the ML and QML methods are their huge computational burden, since the maximization of the log-likelihood involves nonlinear optimization that requires repeated computation of the determinant of matrices of dimension $n \times n$, where n is the size of the data set. Kelejian and Prucha (1999) proposed an alternative estimator for the spatial autoregressive parameter ρ and variance parameter σ^2 based on a generalized moments approach which is computationally simple irrespective of sample size. Further, the conditions required for consistency do not involve the assumption of Gaussian errors. This estimator of ρ is consistent, and can be treated as a nuisance parameter. Then, the asymptotic properties of the regression parameter β based on the estimated ρ can retain desirable properties of OLS for the model where ρ is assumed to be known.

As discussed earlier, there are a growing number of high dimensional applications where there are a large number of potential regressors in X_n, and we need to identify the important covariates. Here, high dimensional data refers to the case where the dimension increases with sample size and ultra-high dimensional the case where the dimensionality grows at a non-polynomial rate as the sample size increases. Typical examples in the spatial regression context include general-to-specific models with indicator saturation and polynomial basis expansions of regressors and their interactions, and spatial Durbin models (Lee and Yu, 2016) where the degree of spatial spillovers in the effect of exogenous covariates are not known *a priori*. We consider an example of the second type in the empirical application developed in this paper. Further, as a result of the wide availability of inexpensive global positioning systems and other technological mechanisms such as Google Trends, the collection of vast quantities of data with geo-referenced sample locations has become possible and the models for spatially correlated data therefore increasingly important (Varian, 2014). Often the number of attributes on which data are collected is so large, often even larger than the sample size, that standard methods of estimation discussed above are rendered infeasible or even impossible. However, the true model still is usually parsimonious, in the sense that most of the potential regressors either have no effect on the response variable Y_n, or at least are uncorrelated with the more important covariates. Thus, the p-dimensional regression parameters are assumed to be sparse with majority of the components being zero.

As discussed before, LASSO and its variants have proved very useful in such sitiations to identify the relevant covariates and estimate their partial effects. Fu and Knight (2000) developed the asymptotic distribution of LASSO-type estimators in the low-dimensional setting where the dimension of regression p is smaller than the sample size n and is fixed. Later in Zhao and Yu (2006), an almost necessary and sufficient Irrepresentable Condition for LASSO is constructed to select the true model consistently both in the fixed p setting and in the large p setting when p can grow with the sample size n. The underlying idea is that the LASSO selects the true model consistently if and (almost) only if the predictors that are not in the true model are "irrepresentable" by predictors that are in the true model. Other results concerning the asymptotic properties of the

LASSO can be found in the Meinshausen and Bühlmann (2006), Bickel et al. (2009) and Bühlmann and van de Geer (2011), among others.

In this paper, we propose a generalized moments LASSO estimator that performs variable selection and estimation simultaneously for the regression parameter β in a two-stage process. Also, we use the consistency property of the estimator for model parameter ρ and the fact that it is a nuisance parameter (Kelejian and Prucha, 1999) to prove that the asymptotic properties of the LASSO estimator of β remain valid even when the model parameter ρ is replaced by its GMM estimator. Both the parameter consistency and model sign consistency of the estimator are addressed. Here, parameter consistency refers to the asymptotic property that, as the sample size n increases to infinity, the resulting sequence of estimates converge in probability to the true parameter value,[3] that is,

$$\hat{\beta}^n \rightarrow_p \beta, \quad as \quad n \rightarrow \infty.$$

An estimator is model sign consistent if and only if the probability that the sign of each component of the estimator equals to that of the true parameter converges to one, that is, there exists $\lambda_n = f(n)$, a function of n and independent of Y_n or X_n, such that

$$\lim_{n\rightarrow\infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1.$$

Based on the framework of Zhao and Yu (2006), we find the appropriate amount of deterministic regularization (penalty) that provides both consistency in model selection, and selection of the true model. Together, we demonstrate superior finite sample performance of our proposed estimator, and its usefulness in applications. Extensions to other spatial contexts and models are also discussed.

[3]Here, and throughout the paper, $\rightarrow_p$ denotes convergence in probability, $\rightarrow_D$ denotes convergence in distribution, and $\rightarrow_{a.s.}$ denotes almost sure convergence.

3 A generalized moments LASSO (GMLASSO) estimator

In this section, we propose a two-stage estimation procedure which combines GMM and LASSO estimation at the same time. We consider a spatial model where the error term is assumed to be spatially autoregressive:

$$\begin{aligned} Y_n &= X_n\beta + u_n, \\ u_n &= \rho M_n u_n + \epsilon_n, \end{aligned} \tag{3.1}$$

where Y_n is the $n \times 1$ vector of observations on the dependent variable, X_n is the $n \times p$ matrix of observations on the explanatory variables, β is the $p \times 1$ vector of unknown model parameters, and u_n is the vector of spatial autoregressive errors with spatial autoregressive parameter ρ, a scalar parameter, M_n is a $n \times n$ spatial weighting matrix of known constants and zero diagonal elements, and ϵ_n is an $n \times 1$ vector of idiosyncratic errors.

For generality, we permit the elements of M_n and ϵ_n to depend on n. We make several standard assumptions as follows:

Assumption 1. *For all n, the idiosyncratic errors $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, generically denoted by ϵ, are independently and identically distributed with zero mean and positive bounded variance σ^2. Additionally, we assume $E(\epsilon^4) < \infty$.*

Assumption 2. *M_n is an exogenous $n \times n$ matrix. All diagonal elements of M_n are zero, $|\rho| < 1$ and the matrix $I - \rho M_n$ is nonsingular for all $|\rho| < 1$.*

M_n is a spatial weights matrix whose elements define the relationship between different units. In a cross-sectional setting, if the ith and jth units are not related, we can set $m_{ij} = m_{ji} = 0$ where m_{ij} is the (i, j)th element of M_n. Often, M_n is set as a contiguity (adjacency) matrix, in which case the non-zero off-diagonal elements are symmetric and have unit value. In other cases, the elements may reflect economic or geographic distances between the units, in which case they are non-negative and symmetric; M_n can still be asymmetric if it is considered in row-standardized form. In other modeling contexts, for

example Bailey et al. (2016), the matrix is assumed to be symmetric but the elements can take values $\{-1, 0, 1\}$. In yet other contexts, the weights can be asymmetric and without any sign or other restrictions, beyond the conditions in Assumption 2; see, for example, Bhattacharjee et al. (2016). In Assumption 2, $|\rho| < 1$ is a stability (spatial granularity) condition, and the invertibility of the matrix $I - \rho M_n$ is to ensure identification in reduced form, that is, the error vector u_n is uniquely defined in terms of the idiosyncratic error vector ϵ_n, as $(I - \rho M_n)^{-1}\epsilon_n$. These assumptions are standard; see for example, Kelejian and Prucha (1999) and Lee (2004).

The first step of our estimation procedure is to obtain a generalized moments estimator of ρ. The estimation procedure follows Kelejian and Prucha (1999), and we outline this below for convenience of exposition. Let $\tilde{u}_n$ be a predictor for the spatially correlated errors u_n. Further, let $\bar{u}_n = M_n u_n$ denote a vector of weighted averages averages, with the weights determined by the rows of M_n (if M_n is row standardized, then the weights sum to unity). If the averaging is applied twice, we have $\bar{\bar{u}}_n = M_n M_n u_n$. Applying the same transformations to the predictors, we have, correspondingly $\tilde{\bar{u}}_n = M_n \tilde{u}_n$ and $\tilde{\bar{\bar{u}}}_n = M_n M_n \tilde{u}_n$. Similarly, let $\bar{\epsilon}_n = M_n \epsilon_n$.

Then, under Assumptions 1 and 2:

$$E[\frac{1}{n}\epsilon_n'\epsilon_n] = \sigma^2 \qquad E[\frac{1}{n}\bar{\epsilon}_n'\bar{\epsilon}_n] = \sigma^2 n^{-1} Tr(M_n' M_n) \qquad E[\frac{1}{n}\bar{\epsilon}_n'\epsilon_n] = 0 \tag{3.2}$$

The spatial autoregressive parameter ρ is included in the above moments equations through the expression $\epsilon_n = u_n - \rho\bar{u}_n$. Thus the equations can be used to obtain a generalized moments estimator for ρ. From Equations (3.1) and (3.2), we obtain

$$\Gamma_n[\rho, \rho^2, \sigma^2]' - \gamma_n = 0. \tag{3.3}$$

Here

$$\Gamma_n = \begin{bmatrix} \frac{2}{n}E(u_n'\bar{u}_n) & \frac{-1}{n}E(\bar{u}_n'\bar{u}_n) & 1 \\ \frac{2}{n}E(\bar{\bar{u}}_n'\bar{u}_n) & \frac{-1}{n}E(\bar{\bar{u}}_n'\bar{\bar{u}}_n) & \frac{1}{n}Tr(M_n'M_n) \\ \frac{1}{n}E(u_n'\bar{\bar{u}}_n + \bar{u}_n'\bar{u}_n) & \frac{-1}{n}E(\bar{u}_n'\bar{\bar{u}}_n) & 0 \end{bmatrix},$$

$$\gamma_n = \begin{bmatrix} \frac{1}{n}E(u_n'u_n) \\ \frac{1}{n}E(\bar{u}_n'\bar{u}_n) \\ \frac{1}{n}E(u_n'\bar{u}_n) \end{bmatrix}$$

Now if we consider the sample moments based on $\tilde{u}_n$, and use these to replace the moments of u_n shown above, similar to Equation (3.3), we get the equation:

$$G_n[\rho, \rho^2, \sigma^2]' - g_n = \nu_n(\rho, \sigma^2), \tag{3.4}$$

where

$$G_n = \begin{bmatrix} \frac{2}{n}\tilde{u}_n'\tilde{\bar{u}}_n & \frac{-1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n & 1 \\ \frac{2}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n & \frac{-1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n & \frac{1}{n}Tr(M_n'M_n) \\ \frac{1}{n}(\tilde{u}_n'\tilde{\bar{u}}_n + \tilde{\bar{u}}_n'\tilde{u}_n) & \frac{-1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n & 0 \end{bmatrix}$$

$$g_n = \begin{bmatrix} \frac{1}{n}\tilde{u}_n'\tilde{u}_n \\ \frac{1}{n}\tilde{\bar{u}}_n'\tilde{\bar{u}}_n) \\ \frac{1}{n}\tilde{u}_n'\tilde{\bar{u}}_n \end{bmatrix}$$

The 3×1 vector $\nu_n(\rho, \sigma^2)$ can be viewed as a vector of residuals, and the GMM estimator for ρ and σ^2 can be defined as the nonlinear least squares estimator, $\hat{\rho}_n$ and $\hat{\sigma}_n^2$, which minimizes the norm of the residual vector. Specifically,

$$(\hat{\rho}_n, \hat{\sigma}_n^2) = \arg\min_{\rho,\sigma^2} \left[G_n[\rho, \rho^2, \sigma^2]' - g_n\right]' \left[G_n[\rho, \rho^2, \sigma^2]' - g_n\right]. \tag{3.5}$$

Several additional assumptions are required to obtain the asymptotic properties of the GMM estimator.

Assumption 3. *The row and column sums of M_n and $(I - \rho M_n)^{-1}$ are bounded uniformly in absolute value. Note that the bound for $(I - \rho M_n)^{-1}$ may depend on ρ.*

Assumption 4. *Let $\tilde{u}_{i,n}$ denote the ith element of $\tilde{u}_n$, we assume that there exist (finite dimensional) random vectors d_{in} and Δ_n such that $|\tilde{u}_{i,n} - u_{i,n}| \leqslant \|d_{in}\|\|\Delta_n\|$ with $n^{-1}\sum_{i=1}^n \|d_{in}\|^{2+\delta} = O_p(1)$ for some $\delta > 0$ and $n^{\frac{1}{2}}\|\Delta_n\| = O_p(1)$.*

Assumption 5. *The smallest eigenvalue of $\Gamma_n'\Gamma_n$ is bounded away from zero, that is, $\lambda_{min}(\Gamma_n'\Gamma_n) \geqslant \lambda_* > 0$, where λ_* may depend on ρ and σ^2.*

For a discussion of these assumptions, we refer to Kelejian and Prucha (1999). Given Assumptions 1 to 5, the nonlinear least squares estimators $\hat{\rho}_n$ and $\hat{\sigma}_n^2$ defined in Equation

(3.5) are consistent estimators of ρ and σ^2, that is, $\hat{\rho}_n \to_p \rho$ and $\hat{\sigma}_n^2 \to_p \sigma^2$ as $n \to \infty$ (Kelejian and Prucha, 1999).

Let us now focus on the context of a spatial regression model whose errors are autoregressive. There are a large number of potential regressors, but β is sparse, that is, most of the elements of β are zero. It is easy to see that, if ρ were known, we could rewrite model (3.1) as

$$(I - \rho M_n)Y_n = (I - \rho M_n)X_n\beta + \epsilon_n.$$

Then, LASSO variable selection and estimation of β can be conducted using the L_1 penalized least squares criterion

$$(Y_n - X_n\beta)'\Sigma_n(\rho)(Y_n - X_n\beta) + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

where $\Sigma_n(\rho) = (I - \rho M_n)'(I - \rho M_n)$; for a given penalty λ_n, we denote this estimator as $\hat{\beta}_L(\rho)$. Of course, in practical applications ρ is typically unknown, and thus the direct LASSO estimator defined above is infeasible. In this case, we may replace ρ by the generalized moments estimator $\hat{\rho}_n$, and propose a feasible generalized moments LASSO (GMLASSO) estimator $\hat{\beta}_L(\hat{\rho}_n)$ for model (3.1) in the second step of the estimation process. To be specific,

$$\hat{\beta}_L(\hat{\rho}_n) = \arg\min_{\beta}(Y_n - X_n\beta)'\Sigma_n(\hat{\rho}_n)(Y - X_n\beta) + \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{3.6}$$

The above function can be numerically optimized using the package "glmnet"[4] in R developed by Friedman et al. (2010). The glmnet algorithms use cyclical coordinate descent, which optimizes the objective function over each parameter successively while keeping others fixed, with the cycles repeating until convergence. The tuning parameter (penalty) λ_n is chosen by cross-validation using a lower bound inferred from the theoretical results discussed below.

[4] An important strength of our methods is that, not only do we allow for correlated errors but also heteroscedastic errors, so long as the error process can be represented by a spatial autoregressive model $u_n = \rho M_n u_n + \epsilon_n$ with homoscedastic idiosyncratic shocks (errors) ϵ_n. And if heteroscedasticity is captured by the spatial model above, then the feasible GMLASSO specification using $\Sigma_n(\hat{\rho}_n)$ (Equation 3.6) applies an appropriate prefiltering that renders standard LASSO as a valid procedure.

4 Asymptotic Properties for fixed p and q

In this section, we consider the asymptotic behavior of the generalized moments LASSO estimator (3.6) in the setting when p (the dimension of all candidate covariates) and q (the dimension of covariates with non-zero coefficients) are both finite and fixed and smaller than the sample size n; that is, $q << p < n$. We show that under the classical setting mentioned above, our proposed GMLASSO estimator achieves consistency in both parameter estimation and model selection.

4.1 Parameter Consistency

In the following theorem, we show that when the tuning parameter λ_n grows at a rate slower than n, the proposed estimator $\hat{\beta}_L(\hat{\rho}_n)$ achieves parameter consistency and if we add more control on the growth rate of λ_n, the asymptotic normality of the estimator can also be established. We make the following assumptions on the regressors.

Assumption 6. *The elements of X_n are nonstochastic and uniformly bounded in absolute value. The matrix $C(\rho) = \lim_{n\to\infty} \frac{1}{n} X_n' \Sigma(\rho) X_n$ is finite and nonsingular for all $|\rho| < 1$ and $\frac{1}{n} \max_{1\leqslant i\leqslant n} z_i z_i' \to 0$, where z_i is the ith row of the matrix $(I - \rho M_n)X_n$.*

Note that (3.1) can be parametrized as the linear model $(I-\rho M_n)Y_n = (I-\rho M_n)X_n\beta+\epsilon_n$. This parametrization is unique if the matrix $C_n = \frac{1}{n} X_n' \Sigma(\rho) X_n$ is nonsingular for all n, and we further assume that $C(\rho)$ is nonsingular. This provides justification behind Assumption 6. The nonstochastic design matrix assumption is strong, and is made here only for expositional simplicity. can be relaxed and is assumed here for explanation simplicity. In fact, the results in this section hold quite generally for random designs. If X_n is a random design matrix, the asymptotic results still apply as long as the conditions in Assumption 6 hold almost surely as $n \to \infty$. Specifically, we can modify Assumption 6 in the following way.

Assumption 6a. *The elements of X_n are potentially stochastic, and satisfy $\frac{1}{n} X_n' \Sigma(\rho) X_n \longrightarrow_{a.s.} C(\rho)$, where $C(\rho)$ is a positive definite matrix for all $|\rho| < 1$. Further,*

$\frac{1}{n}\max_{1\leqslant i\leqslant n} z_i z_i' \longrightarrow_{a.s.} 0$, *where* z_i *is the* i*-th row of the matrix* $(I-\rho M_n)X_n$.

Similar extension can be seen in Zhao and Yu (2006); see also Bühlmann and van de Geer (2011). We view the above assumptions as reasonable. If the x_i are IID with finite second moments, then we can see easily that $\frac{1}{n}X_n'\Sigma(\rho)X_n \longrightarrow_{a.s.} C(\rho)$ and $\max_{1\leqslant i\leqslant n} z_i z_i' = o_p(n)$. Hence conditions in Assumption 6 (and Assumption 6a) are natural. See, however, Lounici (2008) where an argument is made as to why the second part of Assumption 6a can be strong. The assumption can be relaxed further. However, we proceed with this because our focus lies in clarifying what happens to the LASSO in the non-IID setting with a plug-in GMM estimator for the spatial dependence parameter. Further, Assumption 5 requires the design of the relevant covariates to have eigenvalues bounded from below. In the random design case, one simply needs the eigenvalues of the corresponding covariance matrix to be bounded. Then, we have the following result for the GMLASSO estimator (3.6).

Theorem 1. *Under Assumptions 1 to 6, if* $\lambda_n/n \to 0$, *the generalized moments LASSO estimator* $\hat{\beta}_L(\hat{\rho}_n)$ *is consistent, that is,* $\hat{\beta}_L(\hat{\rho}_n) \to_p \beta$ *as* $n \to \infty$. *If we assume further that* $\lambda_n/\sqrt{n} \to \lambda_0 \geqslant 0$, *then*

$$\sqrt{n}(\hat{\beta}_L(\hat{\rho}_n) - \beta) \longrightarrow_D \arg\min(V(w)),$$

where $V(w) = -2w'U + w'C(\rho)w + \lambda_0 \sum_{j=1}^{p}[w_j sgn(\beta_j)I(\beta_j \neq 0) + |w_j|I(\beta_j = 0)]$, *and* $U \sim N(0, \sigma^2 C(\rho))$.

The above theorem establishes parameter consistency of the GMLASSO estimator in the setting where both the dimension of all covariates p and the dimension of non-zero covariates q are fixed and smaller than the sample size n. Further, if we control the rate of convergence of the penalty parameter λ_n in a specific way, the estimator achieves asymptotic normality towards the minimizer of a function $V(w)$. In the function $V(w)$, w is a $p \times 1$ vector, U is a $p \times 1$ random vector with normal distribution, and $C(\rho)$, defined in Assumption 6, involves the spatial parameter ρ and spatial weight matrix M_n. Specifically, if the tuning parameter λ_n grows to infinity at a slower rate than the square root of n, we have asymptotic normality as well. Compared with the asymptotic properties of the traditional (naive) LASSO estimator in the linear model with IID errors,

here we have spatial correlation. We find that the spatial autoregressive parameter ρ is involved in the asymptotic distribution of the GMLASSO estimator and controls the convergence rate; if $\rho = 0$, the asymptotic distribution reduces to the same as that for the traditional LASSO.

4.2 Sign Consistency

Above, we have shown parameter consistency of our generalized moments LASSO (GM-LASSO) estimator $\hat{\beta}_L(\hat{\rho}_n)$. However, a consistent estimator does not necessarily consistently select the correct model. Many applications have a large number of irrelevant predictors, even in the low dimensional settings, and our primary goal is to correctly identify those which are relevant so that the final model will not only fit well but also be parsimonious and easily interpretable. So another property we desire is the model selection consistency of the estimation, which requires that

$$P(\{i : \hat{\beta}_i \neq 0\} = \{i : \beta_i \neq 0\}) \to 1, \quad as \quad n \to \infty.$$

Thus, we follow Zhao and Yu (2006) and achieve the result through sign consistency of the estimator, in which case,

$$sign(\hat{\beta}_L(\hat{\rho}_n)) = sign(\beta),$$

where $sign(\cdot)$ is a function that maps positive elements of the vector argument to 1, negative elements to -1 and zeroes to zero. We denote the above sign consistency condition as

$$\hat{\beta}_L(\hat{\rho}_n) =_s \beta.$$

Note that sign consistency is stronger than model selection consistency, in the sense that, if our estimator is sign consistent, then the model selection consistency condition is automatically satisfied. Further, sign consistency avoids the undesirable situation that the model is estimated only with zeros matched but reversed signs for some of the relevant covariates.

Notation : Assume $\beta = (\beta_1, \cdots, \beta_q, \beta_{q+1}, \cdots, \beta_p)'$ where $\beta_j \neq 0$ for $j = 1, \cdots, q$ and $\beta_j = 0$ for $j = q+1, \cdots, p$. Let $\beta(1) = (\beta_1, \cdots, \beta_q)'$ and $\beta(2) = (\beta_{q+1}, \cdots, \beta_p)'$, and for any p-column matrix Z, write $Z(1)$ and $Z(2)$ as the first q and final $p-q$ columns of Z respectively. Define $C^n(\rho) = \frac{1}{n}[(I - \rho M_n)X_n]'[(I - \rho M_n)X_n]$. By setting

$$\begin{aligned} C^n_{11}(\rho) &= \frac{1}{n}[(I - \rho M_n)X_n](1)'[(I - \rho M_n)X_n](1), \\ C^n_{22}(\rho) &= \frac{1}{n}[(I - \rho M_n)X_n](2)'[(I - \rho M_n)X_n](2), \\ C^n_{12}(\rho) &= \frac{1}{n}[(I - \rho M_n)X_n](1)'[(I - \rho M_n)X_n](2), \end{aligned}$$

and

$$C^n_{21}(\rho) = \frac{1}{n}[(I - \rho M_n)X_n](2)'[(I - \rho M_n)X_n](1),$$

we can express $C^n(\rho)$ as follows:

$$C^n(\rho) = \begin{pmatrix} C^n_{11}(\rho) & C^n_{12}(\rho) \\ C^n_{21}(\rho) & C^n_{22}(\rho) \end{pmatrix}.$$

For the same reason as Assumption 6, here we assume that C^n_{11} is invertible based on the uniqueness of the parametrization of the first relevant q covariates. Since $\hat{\rho}_n$ is a consistent estimator of ρ, the invertibility of $C^n_{11}(\hat{\rho}_n)$ is inherited from that of $C^n_{11}(\rho)$ when the sample size is large enough. The same condition is also valid in the high dimensional case when $p > n$ and Assumption 6 does not hold. Then, this marks the connection with the following Section. In the rest of the paper, we will use the notation C^n to denote $C^n(\hat{\rho}_n)$ unless specified otherwise.

The following proposition places a lower bound on the probability of GMLASSO picking the true model which quantitatively relates to the probability of LASSO selecting the correct model. This is a modification of the Proposition 1 in Zhao and Yu (2006).

Proposition 1. *Assume that* $|C^n_{21}(C^n_{11})^{-1} sign(\beta(1))| \leqslant 1 - \eta$ *holds for some* $\eta > 0$, *where the inequality holds element-wise. Then,*

$$P(\hat{\beta}_L(\hat{\rho}_n; \lambda) =_s \beta) \geqslant P(A_n \cap B_n)$$

for

$$A_n = \left\{ \left|(C^n_{11})^{-1} W^n(1)\right| < \sqrt{n}(|\beta(1)| - \frac{\lambda_n}{2n} \left|(C^n_{11})^{-1} sign(\beta(1))\right| \right\}$$

$$B_n = \left\{ |C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)| \leqslant \frac{\lambda_n}{2\sqrt{n}} \eta \right\}$$

where

$$W^n(1) = \frac{1}{\sqrt{n}} \left[(I - \rho M_n')^{-1} \Sigma(\hat{\rho}_n) \right] (1)' \epsilon_n$$

and

$$W^n(2) = \frac{1}{\sqrt{n}} \left[(I - \rho M_n')^{-1} \Sigma(\hat{\rho}_n) \right] (2)' \epsilon_n$$

In order to prove Proposition 1 and the following Theorem 2, we need the following lemma which is a direct consequence of the Karush-Kuhn-Tucker conditions.

Lemma 1. $\hat{\beta}^n(\lambda) = (\hat{\beta}_1^n, \cdots, \hat{\beta}_j^n, \cdots)$ *are the LASSO estimates defined by*

$$\hat{\beta}^n(\lambda) = \arg\min_{\beta} ||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1$$

if and only if

$$\frac{d||Y_n - X_n\beta||_2^2}{d\beta_j}|_{\beta_j = \hat{\beta}_j^n} = -\lambda sign(\hat{\beta}_j^n) \quad for \quad j \quad such\ that \quad \hat{\beta}_j^n \neq 0$$

$$\left| \frac{d||Y_n - X_n\beta||_2^2}{d\beta_j} \right|_{\beta_j = \hat{\beta}_j^n} \leqslant \lambda \quad for \quad j \quad such\ that \quad \hat{\beta}_j^n = 0.$$

In our context, the generalized moments LASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ is defined to minimize $(Y_n - X_n\beta)'\Sigma(\hat{\rho}_n)(Y_n - X_n\beta) + \lambda_n||\phi||_1$ for some λ_n for all ϕ, where $\hat{\rho}_n$ is the GMM estimator of the parameter ρ in (3.1). Hence, applying Lemma 1 in our case, we have

$$\begin{aligned} \left| \frac{d||(I - \hat{\rho}_n M_n)Y_n - (I - \hat{\rho}_n M_n)X_n\beta||_2^2}{d\beta_j} \right|_{\beta_j = \hat{\beta}_{Lj}(\hat{\rho}_n)} &= -\lambda_n sign(\hat{\beta}_{Lj}(\hat{\rho}_n)) \\ \text{for} \quad j \quad \text{such that} \quad \hat{\beta}_{Lj}(\hat{\rho}_n) &\neq 0 \end{aligned}$$

and

$$\begin{aligned} \left| \frac{d||(I - \hat{\rho}_n M_n)Y_n - (I - \hat{\rho}_n M_n)X_n\beta||_2^2}{d\beta_j} \right|_{\beta_j = \hat{\beta}_{Lj}(\hat{\rho}_n)} &\leqslant \lambda_n \\ \text{for} \quad j \quad \text{such that} \quad \hat{\beta}_{Lj}(\hat{\rho}_n) &= 0, \end{aligned}$$

where $\hat{\beta}_{Lj}(\hat{\rho}_n)$ is the jth element of the estimator $\hat{\beta}_L(\hat{\rho}_n)$. With this result, we are now able to prove the Proposition 1. The proof follows Zhao and Yu (2006) with appropriate

adjustments to our case. Recall that in this section, we focus on the classical setting where q, p,and β are all fixed as $n \to \infty$. Under the above conditions and assumptions, we have the following result about sign consistency of our proposed GMLASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$.

Theorem 2. *For fixed q, p, and β, under Assumptions 1-6, the generalized moments LASSO estimator is sign consistent if the condition*

$$|C_{21}^n (C_{11}^n)^{-1} sign(\beta(1))| \leqslant 1 - \eta$$

holds. That is, for every λ_n that satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \geqslant r$ for any $r > 0$ with $0 \leqslant c < 1$, we have

$$P(\hat{\beta}_L(\hat{\rho}_n) =_s \beta^n) = 1 - o\left(s(\rho) e^{\frac{-n^c}{s^2(\rho)}}\right).$$

From the above result, it is clear that the convergence rate for the estimation method to choose the correct model is a bounded function of the spatial parameter ρ times the exponential of a function of n and s. Here, $s(\rho)$ is the bound for the diagonal elements of $C_{11}^{-1}\sigma^2$ and $C_{22} - C_{21}C_{11}^{-1}C_{12}\sigma^2$. The spatial structure of our model influences the convergence rate. While the convergence rate in the IID case is related only to n, now this depends also on ρ. One remark here is that, for Theorem 2, the effect of the spatial correlation to the estimator in the form of a function of ρ can instead be applied to the penalty parameter λ_n as a lower bound. In this way, additional information can be used for the choice of λ_n besides cross-validation; see also Nandy et al. (2016).

5 Asymptotics for large p and q

In the previous section, we proved parameter consistency and sign consistency, as well as the asymptotic distribution, of our GMLASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ as $n \to \infty$ under the classical setting where p, q, and β are all fixed, and p and q are smaller than n. The setting is simplified in the sense that it is natural to assume the regularity conditions in

Assumption 6:

$$C = \lim_{n\to\infty} \frac{1}{n} X_n' \Sigma(\rho) X_n$$

where C is finite and nonsingular.

However, in practice, there are many situations where large p and thus q are needed; it can either be larger than the sample size n or increase at some rate as n. In the large p and q case, we allow the dimension of the designs C^n to grow and model parameter β to change as n increases, that is, $p = p_n$ and $q = q_n < n$ and $\beta = \beta_n$. Consequently, the assumptions and regularity conditions in the previous sections are not appropriate since C^n may no longer converge and $\beta = \beta_n$ may change as n grows. Towards this situation, we first prove an oracle inequality for the generalized moments LASSO when the design is non-random; this in turn will imply consistency as well. Then, in the second part of this section, we also prove that with high probability we can correctly select the model in the case that $p > n$.

5.1 Parameter consistency

In this section, we prove that, with an appropriate choice of λ_n, the generalized moments LASSO estimator $\hat{\beta}_L(\hat{\rho}_n)$ obeys the following oracle inequality with a probability that can be made arbitrarily close to unity. That is, for large enough n, the condition

$$\frac{\left\| (I - \hat{\rho}_n M_n) X_n (\hat{\beta} - \beta) \right\|_2^2}{n} + \lambda_n ||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}$$

is satisfied with an arbitrarily large probability. The inequality provides a bound for $||\hat{\beta} - \beta||_1$, and thus the estimator is consistent if the bound converges to zero. Here, β is the true value of the unknown parameter, $\lambda_n = O\left(\frac{\log 2p}{n}\right)$, we denote the GMLASSO estimator by $\hat{\beta}$ for notational simplicity, s_0 is the cardinality of the set of nonzero components of β, and S_0 and ϕ_0 are constants depending on the design matrix X_n.

By the definition of the generalized moments estimator:

$$\hat{\beta} := \arg\min_{\phi} \left\{ (Y_n - X_n\phi)' \Sigma(\hat{\rho}_n)(Y_n - X_n\phi) + \lambda_n ||\phi||_1 \right\}.$$

Since $\hat{\beta}$ provides the minima of this penalized objective function, we have the inequality below with change of scale of λ_n:

$$\frac{\left\|(I-\hat{\rho}_n M_n)(Y_n - X_n\hat{\beta})\right\|_2^{\,2}}{n} + \lambda_n||\hat{\beta}||_1 \leqslant \frac{\|(I-\hat{\rho}_n M_n)(Y_n - X_n\beta)\|_2^{\,2}}{n} + \lambda_n||\beta||_1$$

Rearranging terms and using the triangle inequality, we obtain our Basic Inequality:

$$\frac{\left\|(I-\hat{\rho}_n M_n)X_n(\hat{\beta}-\beta)\right\|_2^{\,2}}{n} + \lambda_n||\hat{\beta}||_1 \leqslant \frac{2\epsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}-\beta)}{n} + \lambda_n||\beta||_1. \tag{5.7}$$

Note that the first term on the RHS of the Basic Inequality (5.7) can be easily bounded in terms of the L_1-norm of parameters involved:

$$2\left|\epsilon_n'(I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n(\hat{\beta}-\beta)\right| \leqslant \left(\max_{1\leqslant j\leqslant p} 2|\epsilon_n' T^{(j)}|\right)||\hat{\beta}-\beta||_1$$

where $T^{(j)}$ is the jth column of the matrix $T = (I-\rho M_n')^{-1}\Sigma(\hat{\rho}_n)X_n$.

Next, we introduce the set

$$\Im := \left\{\max_{1\leqslant j\leqslant p} 2|\epsilon_n' T^{(j)}|/n \leqslant \lambda_0\right\}$$

where we arbitrarily assume that $2\lambda_0 \leqslant \lambda_n$ to make sure that on $\Im$ we can get rid of the random part of the problem. Now, we have the following result.

Proposition 2. *Suppose Assumptions 1-5 hold, and further assume all the elements of X_n are nonstochastic and uniformly bounded in absolute value. Then, for all $t > 0$, if we define*

$$\lambda_0 = 2\sigma(\exp[t^2/2]+1)\sqrt{log2p/n},$$

we have

$$P(\Im) \geqslant 1 - K\exp[-t^2/2].$$

for some positive constant K.

The assumptions are inherited from Kelejian and Prucha (1999), from which we draw our GMM estimator $\hat{\rho}_n$. Note that the assumption of nonstochastic and uniformly bounded

X_n is standard in the spatial econometrics and LASSO literatures; see, for example, Kelejian and Prucha (1999), Lee (2004) and Belloni et al. (2012). In our context, this can be replaced by stochastic regressors with strict exogeneity or finite moment conditions; see also Lee (2004) and Su and Yang (2015). However, this would induce substantial analytical complexity, and hence is left outside the scope of the current paper. Likewise, our framework provides a natural setting to extend large dimensional instrument selection (Belloni et al., 2012, 2016) to a dependent data and stochastic regressor context. This is also retained for future work. The proof is given in the Appendix. Since we are in a situation where p is growing with n, and possibly $p > n$, we generally consider the fact that only a few, say s_0, of the β_j are non-zero. To quantify the sparsity of the true β^0, we denote

$$S_0 := \{j : \ \beta_j^0 \neq 0\},$$

so that $s_0 = |S_0|$. In the literature, S_0 is called the active set, and s_0 the sparsity index of β^0.

Before we state the final oracle inequality, using $\lambda_n \geqslant 2\lambda_0$ and the Basic Inequality (5.7), we have on $\Im$,

$$2 \left\| (I - \hat{\rho}_n M_n) X_n (\hat{\beta} - \beta) \right\|_2^2 / n + 2\lambda_n ||\hat{\beta}||_1 \leqslant \lambda_n ||\hat{\beta} - \beta^0||_1 + 2\lambda_n ||\beta||_1.$$

Since

$$||\hat{\beta}||_1 = ||\hat{\beta}_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1 \geqslant ||\beta_{S_0}||_1 - ||\hat{\beta}_{S_0} - \beta_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1,$$

and also,

$$||\hat{\beta} - \beta^0||_1 = ||\hat{\beta}_{S_0} - \beta_{S_0}||_1 + ||\hat{\beta}_{S_0^c}||_1.$$

Therefore,

$$2 \left\| (I - \hat{\rho}_n M_n) X_n (\hat{\beta} - \beta) \right\|_2^2 / n + \lambda_n ||\hat{\beta}_{S_0^c}||_1 \leqslant 3\lambda_n ||\hat{\beta}_{S_0} - \beta_{S_0}||_1. \tag{5.8}$$

Here, λ_n may be viewed as a regularization parameter satisfying the relationship with λ_0 defined in Proposition 2. From Assumption 1, we have $0 < \sigma^2 < b$, hence $\lambda_n = 2\sqrt{b}(\exp[t^2/2] + 1)\sqrt{\frac{\log 2p}{n}}$ is a possible choice.

In order to prove the oracle inequality mentioned at the beginning of this section, we need one more condition on the design matrix corresponding with the consistent estimator of ρ; this is similar to the "compatibility condition" in Bühlmann and van de Geer (2001) with only minor changes. Since we know from inequality (5.8), on $\Im$,

$$||\hat{\beta}_{S_0^c}||_1 \leqslant 3||\hat{\beta}_{S_0} - \beta_{S_0}||_1,$$

we will only require the condition restricted on β_{S_0}. Thus, the compatibility condition in our case is stated as follows.

Condition 1. *Condition 1 is said to be satisfied for the set S_0, if for some constant $\phi_0 > 0$, and for all β satisfying $||\beta_{S_0^c}||_1 \leqslant 3||\beta_{S_0}||_1$, it holds that*

$$||\beta_{S_0}||_1^2 \leqslant (\beta' X_n' \Sigma(\hat{\rho}_n) X_n \beta)\, s_0 / \left(n\phi_0^2\right).$$

Note that, when we obtain a LASSO estimator in the second step of our estimation process, $\hat{\rho}_n$ is considered a known parameter. Finally, we obtain parameter consistency as follows.

Theorem 3. *Suppose Condition 1 holds for S_0, for some $t > 0$, and let the regularization parameter $\lambda_n \geqslant 2\lambda_0$, then on $\Im$, we have*

$$\frac{\left\|(I - \hat{\rho}_n M_n) X_n (\hat{\beta} - \beta)\right\|_2^2}{n} + \lambda_n ||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}.$$

The result also means that with probability at least $1 - K\exp[-t^2/2]$, we have

$$\frac{\left\|(I - \hat{\rho}_n M_n) X_n (\hat{\beta} - \beta)\right\|_2^2}{n} + \lambda_n ||\hat{\beta} - \beta||_1 \leqslant \frac{4\lambda_n^2 s_0}{\phi_0^2}.$$

As discussed earlier, the above result states that with high probability, the L_1 norm of the difference between the estimator and the true value of the parameter of interest is bounded by a function of λ_n and s_0 (same as the dimension of non-zero parameters q_n). Further, the consistency of the estimator is achieved when the bound converges to 0 as $n \to \infty$, and p_n and q_n in this case need to satisfy:

$$\frac{q_n^2 \log 2p_n}{n} \to 0.$$

5.2 Sign Consistency

In section 3.2, we proved sign consistency which infers the model selection consistency of our generalized moments LASSO estimator with a condition similar to the Strong Irrepresentable Condition in Zhao and Yu (2006). Now, we extend the result to sign consistency of the estimator in the high dimensional case when p and q are large and growing with n, following the previous arguments but with an additional assumption:

Assumption 7. *There exists* $0 \leqslant c_1 < c_2 \leqslant 1$ *and* $K_1, K_2, K_3, K_4 > 0$ *so that the following holds:*

$$\begin{aligned} \frac{1}{n}(X_n'\Sigma(\rho))_{ii} &\leqslant K_1, \quad \text{for all } \ i, \\ \alpha' C_{11}^n(\rho)\alpha &\geqslant K_2, \quad \text{for all } \ ||\alpha||_2{}^2 = 1, \\ q_n &= O(n^{2c_1}), \\ n^{\frac{1-c_2}{2}} \min_{i=1,\cdots,q} |\beta_i^n| &\geqslant K_3. \end{aligned} \tag{5.9}$$

Thus, in the case of stochastic design, two conditions need to be satisfied. The first condition in Assumption 7 is straighforward with normalized covariates if the covariates have finite variances, which we assume. The second condition and Assumption 5 require that the design of relevant covariates have eigenvalues bounded from below so that $C_{11}^n(\rho)^{-1}$ is well behaved. For a random design, this usually holds under the sparsity condition in the third equation (Bai, 1999); see also Zhao and Yu (2006). The final condition is the main assumption, requiring a n^{c_2} gap between the decay rate of β^n and $n^{-1/2}$. Under the above assumptions, we can have the following result.

Theorem 4. *Under Assumptions 1-5 and 7, if the condition*

$$\left|(C_{21}^n)\,(C_{11}^n)^{-1}\, sign(\beta(1))\right| \leqslant 1 - \eta$$

holds for some $\eta > 0$*, then for* $p_n = o(n^{2(c_2-c_1)})$*, and for all* λ_n *that satisfies* $\frac{\lambda_n}{\sqrt{n}} = O(n^{\frac{c_2-c_1}{2}})$*, we have*

$$P(\hat{\beta}_L(\hat{\rho}_n; \lambda_n) =_s \beta) \geqslant 1 - O\left(r(\rho)n^{2c_2-2}\right) - o(1) \to 1, \quad as \ \ n \to \infty.$$

Here, we denote $r(\rho)$ as a function of the spatial parameter ρ which controls the maximum of the absolute value of the element in the matrix $(C_{11}^n(\rho))^{-1}$. The term $r(\rho)$ controlling the convergence rate of the estimator to correctly select the true model in our spatial autoregressive errors setting differs from that in the traditional independent data linear regression setting.

6 Simulation Studies

In this section we study the finite sample performance of the generalized moments LASSO estimator (GMLASSO) $\hat{\beta}_L(\hat{\rho}_n)$ in both the low-dimensional setting and the high-dimensional setting and compare these with the traditional LASSO estimator $\hat{\beta}_L$ as well as the ordinary least squares estimator $\hat{\beta}_{OLS}$ (when applicable in the low dimensional case); both the $\hat{\beta}_L$ and the $\hat{\beta}_{OLS}$ ignore spatial dependence in the data. For this purpose, we conduct a two-part Monte Carlo study. Throughout, we set the distribution of ϵ to be Gaussian, and without loss of generality, standard normal $N(0,1)$. This is because the estimators for ρ defined earlier do not depend on σ^2. We consider 6 choices of ρ, covering the range from -1 to 1, together with 5 choices of the sample size n, and thus we have a total of 30 cases for our simulation study. For each case, the results are summarized over 200 Monte Carlo replications.

The design of the study is as follows. Following Kelejian and Prucha (1999), the weight matrix M_n is defined as an idealized weighting matrix M_n, so that each element of u_{ni} is directly related to the one immediately before and after it. We assume the above relationship to be circular, so that u_{nn} is related to u_{n1} and $u_{n,n-1}$, for instance. For simplicity, we specify M_n such that all the non-zero elements of M_n are equal and that the respective rows sum to 1. Our main object of interest lies in the ability of the generalized moments LASSO estimator to consistently choose the correct parameters. The simulation results show the mean (over 200 replicates) of Correctly, Falsely, and Sign-correctly identified components of the parameter for our GMLASSO, traditional LASSO and OLS (only in the low dimensional case), respectively. We also choose a lower

bound for the cross-validation selection of λ_n, which satisfies the conditions implied by our theoretical results.

In the low-dimensional setting, the dimension of the parameter β is chosen as $p = 50$ with the first $q = 5$ non-zero components independently generated from a uniform distribution over the interval $(-2, 5)$ and the rest are zero coefficients. The covariates X_i's are IID from a 50-dimensional Gaussian distribution with each component having mean zero and variance 1. The pairwise correlation is set to $cor(x_{ij}, x_{ik}) = 0.5^{|j-k|}$. The results for this low-dimensional setting are shown in Tables 1 and 2. In each of the tables, the reported figures are the means of the statistics from 200 repetitions; TP (true positive) represents the correctly selected components, FP (false positive) represents the incorrectly selected components and SC (sign correct) represents the the number of correctly selected components with correct sign.

For the high-dimensional case, we set the dimension of the parameter $p = 1000$ but the true number of components that are non-zero is only $q = 20$. M_n is specified in exactly the same way as the low-dimensional setting. Similarly, the first 20 non-zero components are generated independently from a uniform distribution over the interval $(-2, 5)$. The design matrix X_n likewise is the same as that of the low-dimensional design with only change of dimensions. Traditional OLS becomes impossible so we only compare the performance of the generalized LASSO and the traditional LASSO. Note here that, in the traditional LASSO approach, we ignore the autocorrelation of the error u_n and treat the errors as IID. The LASSO estimator is computed using the package "*glmnet*" in R and the penalty parameter λ_n is chosen by 10-fold cross-validation. Another feature that distinguishes our method from the traditional LASSO is the use of a lower bound for λ_n, where this lower bound is motivated by the rates of convergence in our consistency results. The performance of the estimates are reported in Tables 3 and 4.

From the simulation results, we can see clearly that in all of the cases, all the methods considered above can consistently select the non-zero components of the regression parameter when the sample size n becomes larger. What distinguishes the methods is their ability to correctly identify the irrelevant components and set these to zero. From Tables

1 and 2, in the low-dimensional case, it is clear that the traditional LASSO is not suitable for dependent data and OLS works resonably well for all choices of n. However, even though our generalized moments LASSO estimator (GMLASSO) falsely selects more zero components in small sample sizes, the results get much better with increasing data and GMLASSO performs better than the OLS when n exceeds 400. These results are consistent for all choices of the autoregressive parameter ρ.

The above finding of false selection in small samples is not surprising. The traditional LASSO often selects excess covariates even in the IID case. In the literature, the post-LASSO estimator (Belloni and Chernozhukov, 2013) has been proposed to address this issue. In this approach, variables are selected by LASSO first, and a restricted model with only these variables is then estimated by OLS. It appears that the GMLASSO also has a similar issue in small samples. Note that, this false selection is only a small sample problem and vanishes as sample size becomes larger. For this reason, in our application, we use GMLASSO for model selection and post-GMLASSO for obtaining our final estimates.

In the high-dimensional setting, since OLS becomes unavailable, we only compare the performance of the traditional LASSO and our two-stage GMLASSO estimator. We find that, for different choices of ρ, even though both methods can select the non-zero regression coefficients correctly in the main, the performance of the traditional LASSO is poor relative to the generalized moments LASSO estimator in correctly identifying the zero elements. The most interesting observation is the way the LASSO over-selects in the presence of even a little bit of spatial dependence. This inability is not a finite sample bias: if anything, the problem worsens with sample size. There is also an important asymmetry between positive and negative dependence, which has to do with the common finding that inferences are more challenging in negative autocorrelation situations. In applications, often there is negative autocorrelation at finer spatial scales because of spatial competition at the local level; see, for example, Bhattacharjee et al. (2016). However, this negative autocorrelation vanishes at more aggregate spatial scales because negatives convert to positives on compounding. This renders inference on negative au-

tocorrelation at local spatial scales more challenging. In summary, the LASSO loses its selection ability when the errors are not independent.

7 Application to a hedonic house price model

In this section, we illustrate the proposed two-step GMLASSO by application to housing market data for the municipalities of Aveiro and Ílhavo and the adjoining peri-urban and rural area in central Portugal (Figure 1). Specifically, the blue area denotes inland water bodies and are not inhabited. To the north is a large lagoon (Ria de Aveiro), and the Atlantic Ocean occupies the western margin of the plot. Two branches of water extend from the lagoon, and separate the centre from the coastal areas. The western branch extends southwards and westwards to the Atlantic Ocean, trisecting the landmass into three parts. There is limited connection of the beach areas (the strips of land on the western margin) to the mainland. The northern beach is also connected to the mainland by a road which circumvents the lagoon to the north. The other southward extension of the lagoon (towards the centre in Figure 1) is a canal traversed by many bridges; see Bhattacharjee et al. (2016) for the data and for further information. Here, we present estimates of a spatial Durbin hedonic house price model using the post-GMLASSO, that is, model selection by GMLASSO followed by estimation of the selected model by OLS.

The dataset, provided by the firm Janela Digital S.A, relates to the largest real estate advertisement portal in Portugal, and contains $n = 12,467$ observations (houses on sale) sampled from 76 different locations within the above housing market over the period October 2000 and March 2010. Because of lack of sufficient data, we aggregated some of the above locations to create a collection of 71 locations, which are then used in estimation. We estimate a spatial hedonic house price model that allows for location fixed effects, the spatially varying implicit price of living space modeled by the living space elasticity of house price, impact of neighboring living space on own price, and several controls: specifically, (logarithm of) time on the market and 5 statistical factors extracted from data on a large number of hedonic characteristics. Location fixed effects are included to cap-

ture inherent differences across different locations; these may be viewed as neighborhood effects. In addition, there are also location specific housing shocks, and in the context of the spatial model considered here, spatial autocorrelation reflects spillovers of these shocks. Hence, for structural interpretation, we consider accounting for both location fixed effects and spatially correlated errors as being important. Then, the illustrative part of the application relates to estimation of spillover elasticity of living space in neighboring locations to house prices in an index location. This estimation is conducted in a context where the spatial structure (of adjacency or neighborhood membership) is not assumed *a priori*. Thus, for an observation (house) at an index location, we include as potential covariates living space for the respective house, plus proxies for corresponding living spaces at each of the other 70 locations. Specifically, we estimate living space and neighboring living space elasticities by regressing the logarithm of house price per square meter of living area on the logarithm of square meters of living space for the same house and living space in comparable houses in the other locations. This is an example of a spatial hedonic house price model, specifically a spatial Durbin model; see Bhattacharjee at al. (2016) and Lee and Yu (2016) for further discussion and Nowak and Smith (2017) for the use of LASSO for model selection in a different real estate application.

The dataset does not contain information on transaction prices, but only listing prices, which is used as the proxy for price. However, we include as a regressor (logarithm of) time on the market to account for the wedge between listing and sale prices. Location fixed effects are also included to account for unobserved spatial heterogeneity. Potentially, several other hedonic regressors relating to the attributes of the house, as well as the characteristics of the neighborhood, also affect house prices and hence need to be included as controls. The hedonic characteristics (18 physical attributes of each house and 24 location attributes) are subjected to statistical factor analysis to extract 5 factors; see Bhattacharjee et al. (2012) for further details. Thus, apart from location fixed effects, (logarithm of) living space and spillovers, the following six covariates (denoted z) are included and assumed to have spatially fixed coefficients: (logarithm of) time on the market (to account for the wedge between listing and sale prices); factor 1 (access to the centre or to central amenities); factor 2 (access to local services and amenities – health

centres, parks/gardens, etc.); factor 3 (access to beaches, schools and local commerce); factor 4 (physical attributes of the house); and factor 5 (additional house facilities – garage, balcony, central heating, etc.).

The data exhibit substantial spatial effects, which we model quite fully. This is done in three ways. First, we allow full spatial heterogeneity by allowing the shadow price of living space (β_{ii}) to vary across the $L = 71$ locations. In addition, we allow for L location specific fixed effects (α_i) to account for neighborhood level unobserved heterogeneity. Second, we model spatial spillovers in house price shocks by spatial autoregressive errors, where the spatial weights matrix (M_n) is a row-standardized version of inverse geographical distance weights. That is, we first construct a weights matrix where, corresponding to two houses in different locations, the off diagonal elements are reciprocal of the Euclidean distance between the locations; if the houses are in the same location, the corresponding spatial weight is the reciprocal of half the distance of that location to its nearest neighbor location. This weights matrix is then row-standardized by dividing each element by the sum of all entries in its row, and this transformed matrix then constitutes our spatial weights matrix M_n. Third, and most importantly in the context of this work, we allow spillovers of the quality of housing stock from neighboring locations to affect housing price in an index location. The most popular way to accommodate such spillovers in exogenous covariates is the spatial Durbin model (Lee and Yu, 2016):

$$\begin{aligned} Y_n &= X_n\beta + W_nX_n\gamma + u_n, \\ u_n &= \rho M_n u_n + \epsilon_n. \end{aligned}$$

Here, in addition to the (direct) effect of the covariates and the spatial autoregressive errors, there is also the effect coming arising from covariate values in the neighborhood, and captured through a spatial lag term (W_nX_n) with corresponding effect γ. The above spatial Durbin model can have the structural interpretation of capturing the true spillovers in the effect of characteristics in the neighborhood, but may also sometimes be seen as reflection in the reduced form of omitted or inappropriately modeled spatial dependence. It can also be interpreted as accounting for a latent factor structure, either through common correlated effects (Chudik et al., 2016) or statistical factor analysis (Ando and Bai, 2016). Whichever the mechanism that generates such spatial spillovers

in the effects of the regressor, the spatial Durbin model is an important workhorse model in contemporary spatial econometrics.

Typically, the spatial (Durbin) weights matrix W_n is assumed known *a priori*, and usually taken to be the same as M_n. However, mismeasured spatial weights can have serious implications on the inferences drawn, and a current branch of the literature focuses on inferences on the spatial weights themselves; see, for example, Bhattacharjee et al. (2016). Here, we use the GMLASSO for identifying the neighbors that matter and for estimating the implied weights matrix γW_n, which has $L(L-1)$ elements. This allows spillovers and their strength to vary over the spatial domain, which is natural in the current context of hedonic pricing.

In a typical application, this would imply adding covariates for all locations on the right hand side of the regression model and then use LASSO based model selection to estimate both the spatially varying slope (β) and spillovers from other locations (γW_n). In the context of our application, the estimation of a three dimensional functional surface of the spatially varying effect of living space can be tailored to the regression of a linear combination of the effect of living space over nearby locations, besides the effect of living space within each specific location. Thus, the generalized moments LASSO variable selection and estimation method proposed is useful when we select neighboring locations whose living space have an effect on the index location and to estimate how large the effect is; in the process, we can build a parsimonious model by eliminating those locations that are irrelevant for housing prices at each index location.

Importantly, our data has replications that are not in the nature of balanced panel data. Thus, for every house (k) in an index location (i), a corresponding house in any specific other location ($j \in \{1, 2, \ldots, L\}, j \neq i$) is not clearly identifiable. However, note that, if we had panel data with cross-section index $i = 1, \ldots, L$ and replication (time) index $t = 1, \ldots, T$, it would be natural to select the corresponding observation in the other location corresponding to the same replication index; that is, observation (i, t) in location i corresponds to observation (j, t) in location j. This choice is natural because there is an underlying factor structure with a time-specific factor, and the strategy is to

match observations for different cross-section units by this factor (Bhattacharjee et al., 2012). In this case, there are no time replications, but one can match houses in different locations by the underlying factor structure, which is modeled by time on the market and the 5 statistical factors (z). Then, corresponding to the spatial Durbin regressor x_{ik}, we place the corresponding value in an alternate location j at $\widehat{x}_{j,ik} = \widehat{G}_j^x{}^{-1} \widehat{F}_i^z(z_{ik})$, where $\widehat{G}_j^x$ is an estimator of the distribution function of the regressor x in location j, and $\widehat{F}_i^z$ denotes an estimator of the distribution function of the factors (other covariates) z in location i . In the application, we use the empirical distribution function, but any alternate estimator can also be used.[5] Then, our linear model can be described as:

$$y_{ik} = \alpha_i + x_{ik}\beta_{ii} + \sum_{j \neq i} \widehat{x}_{j,ik}\beta_{ij} + z_{ik}{}'\gamma + u_{ik}, \;\; i = 1, \ldots, L, \;\; k = 1, \ldots, n_i, \;\; n = \sum_i n_i.$$

Here y_{ik} is the logarithm of house price per square meter of living space for the k-th replication at the i-th location, while x_{ik} denotes the logarithm of living space, and the corresponding logarithm of living space at each of other locations j is placed at $\widehat{x}_{j,ik}$. Further, u_{ik} is a spatial autoregressive error with spatial weight matrix defined based on the distance between the locations (M_n).

There is one further specific feature in our application. We wish to retain the fixed effects (α_i), location specific elasticities (β_{ii}) and the effects of the additional covariates (γ) in our estimation, and apply model selection only to the spatial Durbin part of the model (β_{ij}). Hence, we first partial the additional regressors out of both the dependent variable (y) and leading covariate (x) and then use GMLASSO and traditional LASSO to the spatial Durbin part. This simplifies model selection to $p = L(L-1) = 4970$ variables with $n = 12467$ observations. Thus, we are in the low dimensional $q < p < n$ setting, but the p is considerably large. Nevertheless, having modeled spatial heterogeneity and error dependence quite fully, we expect the spatial Durbin weights matrix W_n to be very sparse.

Specifically, we wish to identify those locations where there are spillover effects of "living space" from neighboring locations. Hence we implement the two step method for vari-

[5] This approach is related to nonparametric instrumental variables, and proxies based on standard estimators in that literature can also be used; see, for example, Hall and Horowitz (2005).

able selection and estimation for the proposed model. We compare these results with the traditional LASSO method. Table 5 illustrates, based on the first 9 locations in alphabetical order, the differences in number of selected neighbors with spillover effects of living space identified by the generalized moments LASSO (GMLASSO) estimator and the traditional LASSO method. Coinciding with the simulation results, the traditional LASSO estimator substantially over-selects irrelevant variables compared to the GMLASSO and thus offers weak selection power. The GMLASSO selects a parsimonious spatial Durbin model (with only 64 out of a possible 4970 ordered pairs selected) where each location has on average 0.9 neighbors and a median of 0 neighbors. By contrast, the traditional LASSO selects enormously large models with a median of about 40 neighbors.

Finally, we estimate by post-GMLASSO a spatial Durbin hedonic house price model with location fixed effects, spatially varying implicit price of living space, spatial Durbin spillover effects of living space and the effect of 6 additional regressors. These estimates are reported in Table 6, where we only report estimates that are statistically significant at the 5 percent level; the full results are reported in Supplementary material. By post-GMLASSO we mean an OLS estimator based on a sparse model selected by GMLASSO, in the same way that Belloni and Chernozhukov (2013) proposed the post-LASSO as the least squares estimator of a model selected by LASSO. Note that, post-GMLASSO does not explicitly account for the correlation in error terms in (3.1). However, the estimator is still consistent. We chose the post-GMLASSO together with Huber-White corrected heteroscedasticity and autocorrelation consistent standard errors, rather than estimation of a spatial error model, because of its relative simplicity. This final post-GMLASSO model is estimated in the structural form, using OLS, unlike model selection which was conducted in reduced (spatially pre-filtered) form. This is done because of ease of interpretation and substantially reduced computation intensity. Spatial dependence implies that the errors in the structural model (u_n) are spatially correlated and heteroscedastic; hence, we need heteroscedasticity and autocorrelation consistent standard errors.

The estimates offer good interpretation. The effects of the statistical factors and time on the market are significant and have their expected signs. There is substantial unobserved

spatial heterogeneity, as reflected in the wide variation in estimates of location fixed effects. Likewise, the estimated living space elasticities of price shows substantial spatial heteogeneity. Of the 71 locations, 44 have statistically significant estimates, and these estimates lie within the *a priori* expected range of $(0, 1)$. The larger estimated elasticities tend to be at locations closer to the CBD of Aveiro, which is in line with *a priori* expectations. Some estimated elasticities are negative, but none of these are statistically significant; hence we interpret these elasticities as close to zero.

However, our main focus here lies in the estimation of the spatial Durbin part. Here too, some point estimates of spatial Durbin cross-elasticities are very large, but with large standard errors as well; hence, we abstract from interpreting specific point estimates beyond the sign and statistical significance.[6] Of the 64 locations identified by GMLASSO, 13 have estimated spatial Durbin weights that are statistically significant at the 5 percent level. We briefly discuss 8 (of these 13) links that lead towards two locations close to the CBD of Aveiro, Alboi and Beira Mar, which are less than 0.4 km apart; 4 connections lead to each of these locations. Positive and negative externalities are evenly balanced, 2 of each sign for each location. In our context, both positive and negative links have useful interpretation. Positive weights indicate either positive externalities of one location on the other, or perhaps positive externalities generated from the CBD, while negative linkages indicate strong competition between similarly desirable neighborhoods.

The above line of thinking is supported by the evidence. One of the positive influences originates from Agras right at the heart of the city (the CBD), one from Barra which is the main link of the city to the beaches, and the two other locations (Aradas and Quintas)

[6] For example, consider the case of Alboi and Patela. $W(4, 53)$ represents the impact of housing space in Patela upon house price in Alboi, and the estimated elasticity is significantly negative, which reflects high competition between the two locations. In fact, both locations are close to the CBD of Aveiro and competing for residents. Alboi is closer to the CBD, and this closeness is matched by 2.4 times larger floor area (on average) in Patela. It is unlikely that living area in Patela would go much higher, but suppose it were 1% lower (on average), this would imply a 3.5% higher price in Alboi because demand would shift to this neighborhood. However, the point estimate is somewhat imprecise as reflected in a large standard error, and there is a still 0.025 probability that the price is no higher than 1.5%, *ceteris paribas*.

are close by and along main transport links to the city. By contrast, three of the negative links are with Cilhas, Patela and Viso/Caiao, which are emerging neighborhoods with new housing suitable for professionals working in the CBD; hence, locations close to the centre are in severe competition with these neighborhoods. For a similar reason, Avenida Dr. Lourenco Peixinho, a major road with desirable housing close to the centre of Aveiro, also generates negative spillovers on Alboi. Overall, our estimates reflect a parsimonious model that provides good interpretation to the hedonic pricing model with quite full modeling of spatial patterns and dependence. In summary, our proposed GMLASSO method performs well in identifying spatial spillovers, and provides substantial advances over a naive LASSO that clearly does not work with dependent data.

8 Conclusion

With the growth of data rich contexts in business and economics, there has been a proliferation of applications where selection of relevant variables from a large number of candidate covariates is necessary. Various mathods have emerged within the econometrics literature to address this kind of applications. This includes indicator saturation methods (Hendry et al., 2008; Castle and Hendry, 2014), infinite dimensional VARs (Pesaran et al., 2004; Chudik et al., 2016) and LASSO type methods (Caner and Zhang, 2014; Varian, 2014). However, the methods themselves were originally developed for IID settings. Some of these methods have been extended to dependent data settings in a time series context, and to nonstationary latent factor driven spatial strong dependence contexts. More general forms of cross section dependence, such as spatial dependence, has not been addressed in this literature. This is the domain of the current paper.

The literature in spatial econometrics has developed somewhat independently, and this literature has proposed many methods to model spatial (weak) dependence and estimates of such models. This includes likelihood methods (Lee, 2004; Yu et al., 2008) and GMM based methods (Kelejian and Prucha, 1999, 2010; Kapoor et al., 2007; Lin and Lee, 2010). The likelihood based methods, such as maximum likelihood and quasi-maximum

likelihood, are very computation intensive and difficult to apply in large data settings, which has led to the increasing popularity of GMM based methods; see, for example, Brady (2011) and Flores-Lagunes and Schiner (2012). There are also other related methods, such as model averaging (Cuaresma and Feldkircher, 2013) and functional regression (Bhattacharjee et al., 2016). However, there is no specific previous literature on model selection in a spatial context.

This paper considers model selection with a large number of potential regressors, and where the data are spatially dependent through spatial autoregressive errors. We propose generalized moments LASSO (GMLASSO) as a two-stage estimator, combining a GMM estimator to account for spatial dependence with the LASSO for high dimensional variable selection. We derive large sample properties of the estimator, showing in particular parameter consistency and model consistency, as well as asymptotic normality. Our simulation results show that the proposed method dominates the traditional LASSO in moderate and high dimensional settings. In fact, the traditional LASSO provides poor model selection in the presence of even moderate spatial dependence. Importantly, use of the LASSO requires the true model to be sparse. However, it need not be linear model. A wide variety of nonlinear specifications can be considered using transformations of the independent variables, including (as in the case of our application) spatial Durbin variables. The results developed in this paper ensure that the methods will recover the "true" underlying sparse model in large samples. Application to spatial hedonic house price regression in the context of an urban housing market in Portugal highlights the usefulness of the method. Specifically, we highlight spatial error autoregressive dependence and model selection in the context of a spatial Durbin model (Lee and Yu, 2016) with an unknown spillover spatial weights matrix.

Finally, our work suggests several exciting new directions for future research. First, the variable selection result can be used to further update ρ, which results in an iterative process. This can potentially improve efficiency in inference; however, the precise assumptions and statistical results for efficiency will need further research. Second, as discussed earlier, our methods can be taken to the large dimensional instrument selec-

tion context along the lines of Belloni et al. (2012). Whereas they consider an IID setting with nonstochastic regressors and without cross-section dependence, our methods show how one can extend analysis to a context with many stochastic regressors and instruments and with spatial dependence. Third, and likewise, extending the proposed methods to a spatial autoregressive model is a challenging and interesting problem. In this context, one can potentially extend the instrumental variables LASSO and post-LASSO to estimate a spatial weights matrix, following Ahrens and Bhattacharjee (2015) and Lam and Souza (2016). Fourth, consideration of endogenous spatial weights along the lines of Bhattacharjee et al. (2016) may be useful. Fifth, extension of LASSO based variable selection to spatial dependence context in two stages using error covariance estimates, along the lines of Nandy et al. (2016) also holds promise. Finally, extension of the methods to the context of infinite dimensional VARs and indicator saturation will be useful.

References

[1] Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics (MDPI)* **3**(1), 128-155.

[2] Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics* **31**, 163-191.

[3] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic: Boston, MA.

[4] Bai, Z.D. (1999). Methodologies in spectral analysis of large dimensional random matrices: A review. *Statistica Sinica* **9**, 611-677.

[5] Bailey, N., Holly, S. and Pesaran, M.H. (2016). A Two-Stage Approach to Spatio-Temporal Analysis with Strong and Weak Cross-Sectional Dependence. *Journal of Applied Econometrics* **31**, 249-280.

[6] Barry, R. P. and Pace, R.K. (1999). Monte Carlo Estimates of the Log Determinant of Large Sparse Matrices. *Linear Algebra and its Applications* **289**, 41-54.

[7] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica* **80**, 2369-2429.

[8] Belloni, A. and Chernozhukov, V. (2011). High Dimensional Sparse Econometric Models: An Introduction. arXiv:1106.5242v2 .

[9] Belloni, A. and Chernozhukov, V. (2013). Least Squares After Model Selection in High-dimensional Sparse Models *Bernoulli* **19**, 521-547.

[10] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root LASSO: pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791-806.

[11] Belloni, A., Chernozhukov, V. and Wei, Y. (2016). Post-Selection Inference for Generalized Linear Models with Many Controls. *Journal of Business and Economic Statistics*, forthcoming.

[12] Bhattacharjee, A., Castro, E., Maiti, T. and Marques, J. (2016). Endogenous spatial regression and delineation of submarkets: a new framework with application to housing markets. *Journal of Applied Econometrics* **31**, 32-57.

[13] Bhattacharjee, A., Castro, E.A. and Marques, J.L. (2012). Understanding spatial diffusion with factor-based hedonic pricing models: the urban housing market of Aveiro, Portugal. *Spatial Economic Analysis* **7**(1), 133-167.

[14] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics* **37**, 1705-1732.

[15] Brady, R.R. (2011). Measuring the diffusion of housing prices across space and over time. *Journal of Applied Econometrics* **26**(2), 213-231.

[16] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data.* Springer.

[17] Caner, M. and Zhang, H.H. (2014). Adaptive Elastic Net for Generalized Methods of Moments. *Journal of Business and Economic Statistics* **32**(1), 30-47.

[18] Castle, J.L., Doornik, J.A., Hendry, D.F. and Pretis, F. (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics (MDPI)* **3**(2), 240-264.

[19] Castle, J.L. and Hendry, D.F. (2014). Model selection in under-specified equations with breaks. *Journal of Econometrics* **178**, 286-293.

[20] Chudik, A. and Pesaran, M.H. (2011). Infinite-dimensional VARs and factor models. *Journal of Econometrics* **163**(1), 4-22.

[21] Chudik, A., Grossman, V. and Pesaran, M.H. (2016). A multi-country approach to forecasting output growth using PMIs. *Journal of Econometrics*, forthcoming.

[22] Cuaresma, C. and Feldkircher, M. (2013). Spatial filtering, model uncertainty and the speed of income convergence in Europe. *Journal of Applied Econometrics* **28**(4), 720-741.

[23] Fan, J. and Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica* **20**, 101-148.

[24] Feng, W., Lim, C., Maiti, T. and Zhang, Z. (2016). Spatial Regression and Estimation of Disease Risks: A Clustering based Approach. *Statistical Analysis and Data Mining*, forthcoming.

[25] Flores-Lagunes, A. and Schnier, K.E. (2012). Estimation of sample selection models with spatial dependence. *Journal of Applied Econometrics* **27**(2), 173-204.

[26] Friedman, J. Hastie, T. and Tibshirani, R. (2010). Regularization Paths For Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1-22.

[27] Fu, W. and Knight, K. (2000). Asymptotics for LASSO-type estimators. *Annals of Statistics* **28**, 1356-1378.

[28] Geyer, C.J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript.*

[29] Hall, P. and Horowitz, J.L. (2005). Nonparametric Methods for Inference in the Presence of Instrumental Variables. *Annals of Statistics* **33**, 2904-2929.

[30] Hendry, D.F., Johansen, S. and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics* **33**, 317-335. Erratum, 337-339.

[31] Ishwaran, H. and Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**(2), 730-773.

[32] Johansen, S. and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In: *The Methodology and Practice of Econometrics*; Castle, J.L. and Shephard, N. (Eds.), Oxford University Press: Oxford UK, 1-36.

[33] Kapoor, M., Kelejian, H.H. and Prucha, I.R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics* **140**, 97-130.

[34] Kelejian, H. H. and Prucha, I. R. (1999). A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economics Review* **40**, 509-533.

[35] Kelejian, H.H. and Prucha, I.R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* **157**, 53-67.

[36] Kock, A.B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186**(2), 325–344.

[37] Lam, C. and Souza, P. C. (2016). Regularization for spatial panel time series using the adaptive LASSO. *Journal of Regional Science*, forthcoming.

[38] Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**(6), 1899-1925.

[39] Lee, L.-F. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* **154**, 165-185.

[40] Lee, L.-F. and Yu, J. (2016). Identification of spatial Durbin panel models. *Journal of Applied Econometrics* **31**(1), 133-162.

[41] Lin, X. and Lee, L.-F. (2010). GMM estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics* **157**(1), 34-52.

[42] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2**, 90-102.

[43] Meinshausen, N. and Bühlmann, P. (2006). High Dimensional Graphs and Variable Selection with the LASSO. *Annals of Statistics* **34**, 1436-1462.

[44] Nandy, S., Lim, C. and Maiti, T (2016). Additive model building for spatial regression. *Journal of the Royal Statistical Society Series B*, forthcoming.

[45] Nowak, A. and Smith, P. (2017). Textual Analysis in Real Estate. *Journal of Applied Econometrics*, forthcoming.

[46] Ord, J. K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association* **70**, 120-126.

[47] Pesaran, M.H., Schuermann, T. and Weiner, S.M. (2004). Modelling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business and Economic Statistics* **22**(2), 129-162.

[48] Pollard, D. (1991). Asymptotic for least absolute deviation regression estimators. *Econometric Theory* **7**, 186-199.

[49] Smirnov, O. and Anselin, L. (2001). Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach. *Computational Statistics and Data Analysis* **35**, 301-319.

[50] Stock, J.H. and Watson, M.W. (2002). Forecasting using principle components from a large number of predictors. *Journal of the American Statistical Association* **97**, 1167-1179.

[51] Su, L. and Yang, Z. (2015). QML estimation of dynamic panel data models with spatial errors. *Journal of Econometrics* **185**(1), 230-258.

[52] Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society. Series B* **58**, 267-288.

[53] Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives* **28**, 3-28.

[54] Whittle, P. (1954). On stationary Processes in the Plane. *Biometrica* **41**, 434-449.

[55] Yu, J., de Jong, R.M. and Lee, L.-F. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both N and T are large. *Journal of Econometrics* **146**(1), 118-134.

[56] Zhao, P. and Yu, Bin. (2006). On Model Selection Consistency of LASSO. *Journal of Machine Learning Research* **7**, 2541-2563.

[57] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

[58] Zou, H. and Zhang, H. (2009). On the Adaptive Elastic-Net With a Diverging Number of Parameters. *Annals of Statistics* **37**, 1733-1751.

Table 1: Means of TP, FP, SC for 200 Monte Carlo replications

Low dimensional case ($p < n$), Positive autocorrelation

TP–true positive, FP–false positive, SC–sign correct

n		ρ= 0.25			ρ= 0.5			ρ= 0.75		
		TP (5)	FP (45)	SC (5)	TP (5)	FP (45)	SC (5)	TP (5)	FP (45)	SC (5)
100	GMLASSO	5	19.85	5	5	19.87	5	5	19.71	5
	LASSO	5	14.90	5	5	19.24	5	5	24.64	5
	OLS	5	2.93	5	5	3.43	5	4.91	4.72	4.91
200	GMLASSO	5	13.10	5	5	11.96	5	5	9.90	5
	LASSO	5	14.88	5	5	21.03	5	5	25.33	5
	OLS	5	3.46	5	5	4.51	5	5	6.55	5
400	GMLASSO	5	6.61	5	5	6.33	5	5	6.7	5
	LASSO	5	15.73	5	5	20.85	5	5	26.29	5
	OLS	5	3.39	5	5	5.08	5	5	7.3	5
600	GLASS	5	1.66	5	5	1.59	5	5	1.76	5
	LASSO	5	15.78	5	5	20.37	5	5	26.10	5
	OLS	5	3.74	5	5	5.23	5	5	7.23	5
800	GLASS	5	0.23	5	5	0.25	5	5	0.35	5
	LASSO	5	15.11	5	5	21.62	5	5	26.18	5
	OLS	5	3.76	5	5	5.44	5	5	7.46	5

Table 2: Means of TP, FP, SC for 200 Monte Carlo replications

Low dimensional case ($p < n$), Negative autocorrelation

TP–true positive, FP–false positive, SC–sign correct

n		$\rho = -0.25$			$\rho = -0.5$			$\rho = -0.75$		
		TP (5)	FP (45)	SC (5)	TP (5)	FP (45)	SC (5)	TP (5)	FP (45)	SC (5)
100	GMLASS	5	16.2	5	5	14.20	5	5	11.08	5
	LASSO	5	8.38	5	5	5.32	5	5	4.07	5
	OLS	4.99	1.9	4.99	4.93	1.63	4.93	4.64	1.1	5
200	GMLASSO	5	11.87	5	5	11.97	5	5	10.46	5
	LASSO	5	7.3	5	5	4.56	5	5	2.75	5
	OLS	5	1.6	5	5	0.93	5	5	0.6	5
400	GMLASSO	5	2.83	5	5	2.90	5	5	2.53	5
	LASSO	5	6.79	5	5	3.82	5	5	2.54	5
	OLS	5	1.44	5	5	0.66	5	5	0.29	5
600	GMLASSO	5	0.33	5	5	0.37	5	5	0.23	5
	LASSO	5	6.58	5	5	4.06	5	5	2.21	5
	OLS	5	1.27	5	5	0.55	5	5	0.25	5
800	GMLASSO	5	0.01	5	5	0.02	5	5	0.03	5
	LASSO	5	6.21	5	5	3.64	5	5	2.22	5
	OLS	5	1.22	5	5	0.47	5	5	0.22	5

Table 3: Means of TP, FP, SC for 200 Monte Carlo replications

High dimensional case ($p > n$), Positive autocorrelation

TP–true positive, FP–false positive, SC–sign correct

n		ρ= 0.25			ρ= 0.5			ρ= 0.75		
		TP (20)	FP (980)	SC (20)	TP (20)	FP (980)	SC (20)	TP (20)	FP (980)	SC (20)
100	GMLASSO	15.46	77.36	15.44	15.62	79.71	15.61	15.2	80.74	15.2
	LASSO	16.44	105.63	16.40	16.41	113.32	16.39	15.94	117.49	15.93
200	GMLASSO	19.57	96.53	19.57	19.51	104.47	19.51	19.26	122.23	19.26
	LASSO	19.73	116.65	19.73	19.73	168.06	19.72	19.58	254.2	19.58
400	GMLASSO	19.97	93.05	19.97	19.93	116.29	19.93	19.74	158.00	19.74
	LASSO	19.98	129.73	19.98	19.99	194.94	19.99	19.86	277.33	19.86
600	GMLASSO	20	48.51	20	19.98	69.07	19.98	19.9	126.99	19.9
	LASSO	20	140.29	20	19.99	226.32	19.99	19.96	325.65	19.96
800	GMLASSO	20	14.90	20	19.99	23.68	19.99	19.97	61.81	19.97
	LASSO	20	150.89	20	20	258.2	20	19.99	374.55	19.99

Table 4: Means of TP, FP, SC for 200 Monte Carlo replications

High dimensional case ($p > n$), Negative autocorrelation

TP–true positive, FP–false positive, SC–sign correct

n		$\rho = -0.25$			$\rho = -0.5$			$\rho = -0.75$		
		TP (20)	FP (980)	SC (20)	TP (20)	FP (980)	SC (20)	TP (20)	FP (980)	SC (20)
100	GMLASSO	15.09	69.58	15.07	14.35	70.90	14.34	13.93	67.04	13.89
	LASSO	15.98	88.43	15.95	14.92	82.18	14.88	13.64	60.09	13.61
200	GMLASSO	19.62	72.61	19.62	19.56	71.12	19.56	19.50	68.42	19.50
	LASSO	19.65	71.69	19.65	19.46	58.9	19.46	19.05	48.67	19.05
400	GMLASSO	19.97	43.43	19.97	19.99	41.87	19.99	20	40.36	20
	LASSO	19.99	54.26	19.99	19.95	35.72	19.95	19.76	23.83	19.76
600	GMLASSO	20	14.10	20	20	15.27	20	20	15.07	20
	LASSO	20	46.13	20	20	28.38	20	19.95	16.71	19.95
800	GMLASSO	20	3.07	20	20	4.12	20	20	4.2	20
	LASSO	20	42.16	20	20	24.28	20	19.99	14.14	19.99

1a

1.*jpg*

Figure 1: **The Aveiro-Ílhavo housing market**

Table 5: Estimated neighbors with significant spillover effect of living space (selected locations)

Location code and name	Estimated no. of neighbors	
	GMLASSO	LASSO
1 (Agras)	4	56
3 (Alagoas)	0	63
4 (Alboi)	8	56
5 (Aradas)	0	26
6 (Aven Dr. L. Peixinho)	0	25
7 (Azenha de Baixo)	0	8
8 (Azurva)	0	43
9 (Bairro de Santiago)	5	25
10 (Bairro de Liceu)	0	11